Data 620

ASSIGNMENT

[**Assignment**](https://bbhosted.cuny.edu/webapps/assignment/uploadAssignment?content_id=_39523736_1&course_id=_1705333_1&group_id=&mode=view)

In this week's assignment, you are asked to analyze high frequency words.

Please answer the fo llowing questions in an Jupyter Notebook, posted to GitHub.

1. Choose a corpus of interest.
2. How many total unique words are in the corpus?  (Please feel free to define unique words in any interesting, defensible way).
3. Taking the most common words, how many unique words represent half of the total words in the corpus?Identify the 200 highest frequency words in this corpus.
4. Create a graph that shows the relative frequency of these 200 words.
5. Does the observed relative frequency of these words follow Zipf’s law? Explain.
6. In what ways do you think the frequency of the words in this corpus differ from “all words in all corpora.”

Project 3

**This is a Team Project!**  For this project, please work with the entire class as one collaborative group!  Your project should be submitted (as an Jupyter Notebook via GitHub). You should all submit a link to the same repository. 

Using any of the three classifiers described in chapter 6 of Natural Language Processing with Python, and any features you can think of, build the best name gender classifier you can. Begin by splitting the Names Corpus into three subsets: 500 words for the test set, 500 words for the dev-test set, and the remaining 6900 words for the training set. Then, starting with the example name gender classifier, make incremental improvements. Use the dev-test set to check your progress. Once you are satisfied with your classifier, check its final performance on the test set. How does the performance on the test set compare to the performance on the dev-test set? Is this what you'd expect?   
  
Source: Natural Language Processing with Python, exercise 6.10.2.

The ability to be an effective member of a virtual team is highly valued in the data science job market.